[**DSCI 5180 Section 501 - Introduction to the Business Decision Process (Spring 2024 1)**](5180-FINAL%20PROJECT%20DATASET-%20work.xlsx)

**FINAL PROJECT**

**DATA SET NAME**

Cause of Deaths around the World (Historical Data)  
Global Mortality Insights: Historical Dataset on Causes of Death

**DATA SET**

[**5180-FINAL PROJECT DATASET- work.xlsx**](5180-FINAL%20PROJECT%20DATASET-%20work.xlsx)

**DESCRIPTION**

The population's wellbeing is impacted by mortality or thoughts like life hope, which are predicated on estimations of mortality. Incapacity balanced life a long time (DALYs) are a valuable instrument for measuring mortality and predominant maladies, which are critical components in deciding wellbeing results. Considering of this as a conceivable venture point right presently. DALYs are a standardized measurement utilized to degree misplaced wellbeing that empowers coordinate comparisons of malady burdens among populaces, over national borders, and over time.

**DATA SET OVERVIEW**

We have chronicled information on numerous causes of mortality for individuals of all ages around the world in this dataset. In 2019, cardiovascular infections account for the most noteworthy portion of the around the world malady burden. Cancers, infant ailments, musculoskeletal conditions, respiratory contaminations, and mental and sedate manhandle clutters come another. All inclusive, these reasons are positioned in a really distinctive arrange.

**STRUCTURE OF DATASET**

Number of Rows: 6091

Number of Columns: 36

**DATA SET REFERENCE**

Dataset is pulled from Our world in data which is open access data.

**‘’**[**https://www.kaggle.com/datasets/iamsouravbanerjee/cause-of-deaths-around-the-world**](https://www.kaggle.com/datasets/iamsouravbanerjee/cause-of-deaths-around-the-world)**’’**

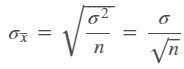
**MODULE 1: Using - The Distribution of the Sample Mean and the Central Limit Theorem**

As per our world in data the mean of cardiovascular diseases for Unite states is 881278 and standard deviation is 35134. What would be the increase in cardiovascular diseases over period of years. Assuming the claim is true for a sample of 20, what is the probability that mean of cardiovascular diseases for United states would differ from the population mean by more than 1000. Rounding the answer to four decimal places.

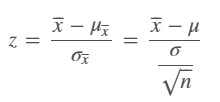
**Solution:**

Since the mean is 881278 which means the value of x̅ should not fall between 880278 and 882278. Also, for the error of estimation to be more than 1000.

From above we see mean(µ) is 881278, standard deviation (σ) is 35134, x̅ is 882278, sample size(n) is 20. Using below formula we get σx̅ as 7856.201232.



In order to find the probability that x̅ is more than 1000 of the given mean, using below formula the Z value is calculated to 0.127288. From the z table we can find:



P (Z > 0.13) which comes to 0.5517. As the condition follows 1−P (μ−1000 < x̅ < μ+1000) the value turns to 0.4483+ 0.4483 = 0.8966.

There is around 89% chance that cardiovascular diseases for United states would differ from the population mean by more than 1000.

**MODULE 2:** **Using - Estimating the Population Mean, Sigma (σ) Unknown**

From a random sample of 25 over the years of tuberculosis disease in Australia. The average tuberculosis disease was 85 with a standard deviation of 7.3. Consider 95% as confidence interval to estimate the average tuberculosis disease.

**Solution:**

From the above question we get x̅ is 85, S is 7.3, α is 0.05 and n is 25.

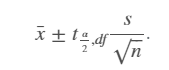
As the sample size is 25 which is less than 30 i.e n<30, we will use the *t*-distribution to construct the confidence interval. The degrees of freedom(df) for above is calculated as

df = n-1 = 25-1 = 24

Using T.INV.2T (probability, degrees of freedom) the *t*-value corresponding to 24 degrees of freedom at 95% confidence (α=0.05) is

=T.INV.2T (0.05,24) = 2.064

Also, we know that formula of confidence interval for the population mean when σ is unknown assuming normally distributed is



After substituting values in above formula, we get the lower end point and upper end point as follows:

**Lower end point: 85+(-2.064) \*1.46 = 81.98671**

**Upper end point: 85+2.064\*1.46 = 88.01329**

**At alpha 0.05 the mean lies between the range 81.98671 and 88.01329.**

**MODULE 3: Using -Testing a Hypothesis about a Population Mean, Sigma Unknown (t-value)**

As per world data the average life expectancy of India for female is 68. Assuming population distribution is normal and the claim is less than reported from a sample of 25, the mean reports 65.9 with standard deviation of 4.2 over period of years. Consider 0.05 level of significance to test the claim, by stating null and hypothesis including test statistic and later interpret and conclude the decision to be taken.

**Solution:**

From above question we have µ as 68, x̅ as 65.9, S as 4.2, n as 25, α is 0.05 and one tail test (left tail).

Degrees of freedom = df = n-1 = 25-1 = 24

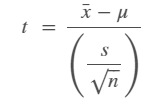
**Stating Null and Alternate hypothesis:**

Null Hypothesis: µ = 68

Alternate Hypothesis: µ < 68

**Finding test statistic:**

We use t-test statistic because the claim is about a population mean and the population is normally distributed with unknown population standard deviation and random sample.



Using above formula and substituting the values as per mentioned, we get t value as -2.500.

**Drawing conclusion and interpreting the decision based on above details:**

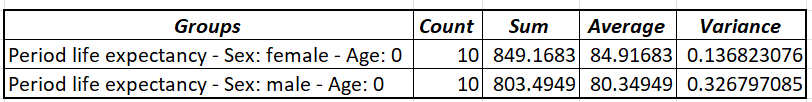
As mentioned in question α is 0.05 and degrees of freedom (df) is 24. Using below formula, p value is 0.009827. Here P value is less than alpha i.e 0.009827 < 0.05 which means we reject null hypothesis.

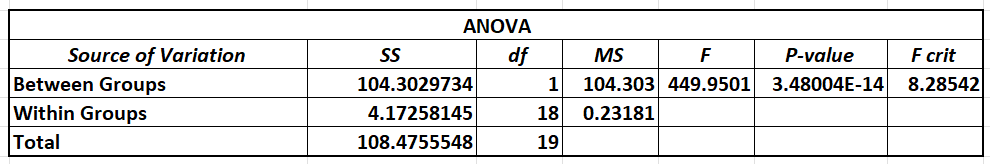
T. DIST.RT (x, df) = T. DIST.RT(-2.500,24) = 0.009827

We reject the null hypothesis and conclude that there is sufficient evidence at 0.05 significance to support the claim that average life expectancy of female in India is 68.

**MODULE 4:** **Using -** **ANOVA Comparison of Multiple Means**

The following information is regarding the life expectancy of male and female of Italy country from 2010 to 2019. Using α=0.01 and the below ANOVA output, provide the conclusion if data provides sufficient evidence about significant difference among average life expectancy is different for both the genders.





**Solution:**

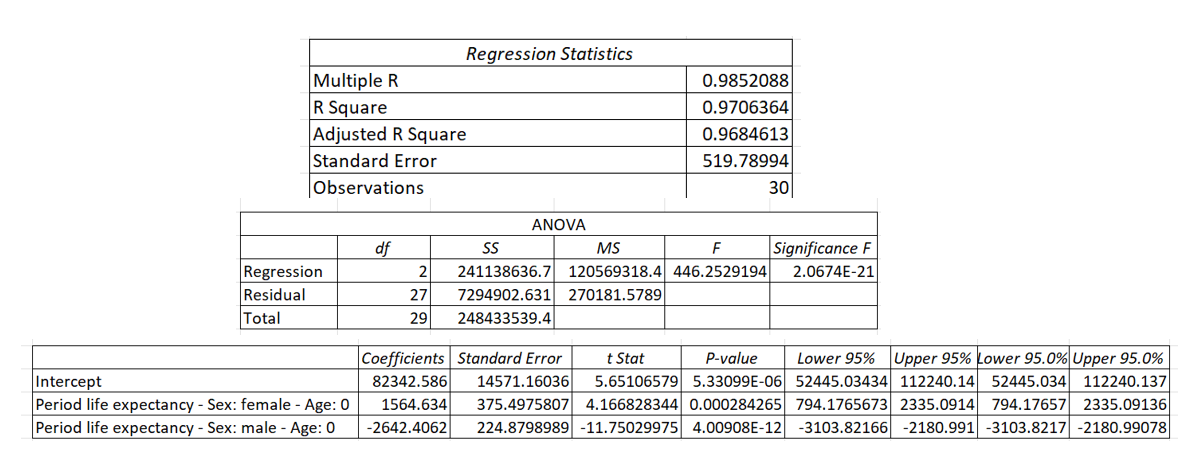
From above ANOVA table, we see F value is 449.9501 and F critical value is 8.28542. From these we can determine F > F critical value and F is significant. When F is significant, we reject null hypothesis with sufficient evidence as per the condition given in question.

**Therefore, we reject the null hypothesis of equal life expectancy for both the genders. At 0.01 there is sufficient evidence that there is difference among the average life expectancies for both male and female genders in Italy.**

**MODULE 5:** **Using -** **Multiple Regression Model**

Consider the following data regarding male and female life expectancy being independent variables and Tuberculosis as dependent variable of Thailand country and the estimated regression equation is as follows:

**Estimated alcohol use = 82342.5858+ 1564.6339 (female life expectancy) -2642.4062 (male life expectancy)**



A female individual with 65 years of life expectancy and a male individual with 72 years of life expectancy. According to this model, what would be the estimated Tuberculosis range?

**Solution:**

We know the general form of estimated regression for two independent variable is



From above question we get x1 as female life expectancy and x2 as male life expectancy and the equation as:

**Estimated alcohol use = 82342.5858+ 1564.6339 (female life expectancy) -2642.4062 (male life expectancy)**

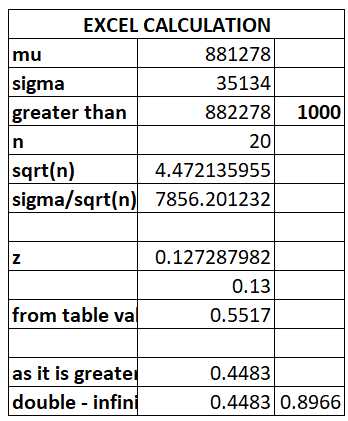
To find the estimated Tuberculosis, substitute the values of male and female life expectancy in estimated regression equation, which gives as -6209.45438. Thus, estimated Tuberculosis range over years when both genders are considered turns to -6209.45438.

**Regression Analysis:**

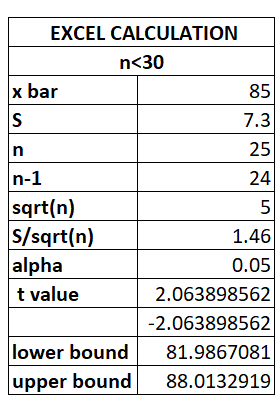
In order to compare p-value with alpha we consider Significance F column. From above we see p-value as 2.0674031952592E-21. Dealing with alpha at 0.05 significance level, we can say at 0.05 level of significance there is sufficient evidence. Also, Coefficient determination(R^2) is 0.97063 which is near to one and Adjusted coefficient determination (adj R^2) is 0.96846. We consider adjusted coefficient determination for multiple regression and in this scenario, we can say that around 96% the regression model and relationship between independent variables and dependent variable is statistically significant.

**APPENDIX:**

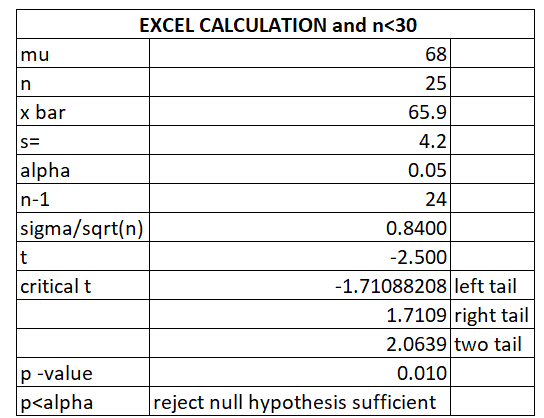
**MODULE 1: The Distribution of the Sample Mean and the Central Limit Theorem**



**MODULE 2: Estimating the Population Mean, Sigma (σ) Unknown**

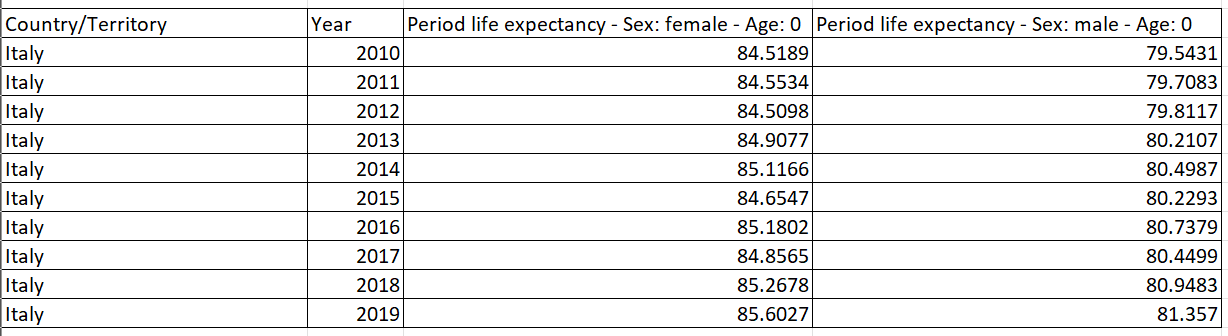
****

**MODULE 3:** **Testing a Hypothesis about a Population Mean, Sigma Unknown (t-value)**



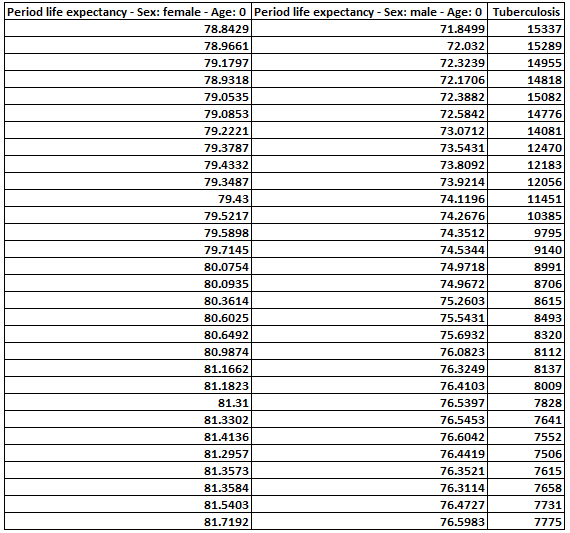
**MODULE 4:** **Using -** **ANOVA Comparison of Multiple Means**

**Data set:**



**MODULE 5:** **Using -** **Multiple Regression Model**

**Data set:**

****

**SAS OUTPUT:**

[**SAS.pdf**](SAS.pdf)